DOCUMENT RESUME

ED 106 359 TM 004 475

AUTHOR Scheuneman, Janice
TITLE A New Method of Assessing Bias in Test Items.
PUB DATE [Apr 75]
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Achievement Tests; Content Analysis; *Cultural Factors; Ethnic Groups; *Item Analysis; Racial Factors; Statistical Analysis; *Test Bias; *Test Construction; Testing; Testing Problems; *Test Validity
IDENTIFIERS Metropolitan Readiness Test

ABSTRACT

In order to screen out items which may be biased against some ethnic group prior to the final selection of items in test construction, a statistical technique for assessing item bias was developed. Based on a theoretical formulation of R. B. Darlington, the method compares the performance of individuals who belong to different ethnic groups, but have equal scores on a subtest containing the item. A chi square technique is used to evaluate differences in performance and is demonstrated on data collected during the item analysis phase of the current revision of the Metropolitan Readiness Tests. (Author)

A NEW METHOD OF ASSESSING BIAS IN TEST ITEMS

Janice Scheuneman
Test Department
Harcourt Brace Jovanovich, Inc.

A paper presented at the annual meeting of the
American Educational Research Association
Washington, D. C., March 31, 1975

A NEW METHOD OF ASSESSING BIAS IN TEST ITEMS

Janice Scheuneman
Harcourt Brace Jovanovich, Inc.

While test bias is defined in many ways and must ultimately be assessed in terms of how a test is to be used, it would seem desirable in the construction of new measuring instruments to screen out items which are likely to be biased before assembling the final forms of the test. In the 1976 revision of the Metropolitan Readiness Tests (MRT) a strong effort was made to eliminate biased items during the item analysis process. Items were reviewed for possible bias in the content by the authors, staff members and minority group consultants, but it was felt a more rigorous statistical procedure was also required.

Where criterion measures are not available, bias is most frequently defined as item by group interaction. That is, groups under consideration may not be equivalent in the ability being measured, but in an unbiased test the differences between them are expected to be consistent across items. Procedures have been developed to determine which items are contributing most to an interaction, if it exists, but the focus is on the test or subtest as a whole rather than on the items. (See for example, Cleary & Hilton, 1968, and Angoff & Ford, 1973.) With a large item pool grouped somewhat arbitrarily into experimental test forms, however, there is little interest in the subtest as a whole. The question in this case is whether an item would be biased when placed into some set of items measuring the same skill. The item-group interaction procedures were not designed to address this question. Therefore, a new technique for assessing item bias was developed.

In this study, an item is considered unbiased if, for persons with the same ability in the area being measured, the probability of a correct response on the item is the same regardless of the population group membership of the individual. Assuming that the subtest score is a reasonably valid measure of the ability in question, this definition can be stated in more operational terms as follows: An item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered. Using this definition and standard statistical techniques, it is possible to determine the probability that any item in the pool is unbiased. Where the probability is sufficiently low, the item is discarded.

Method:

Data were gathered during the fall 1973 item analysis program for the Metropolitan Readiness Tests. The experimental version of the MRT used for item analysis consisted of 14 forms, seven at each of two levels, Kindergarten and beginning Grade 1, with four to six subtests on each form. The Level I tests consisted of 28 subtests in nine areas while Level II included 37 subtests in twelve areas for a total of 65 subtests. The subtests were designed to measure various aspects of visual discrimination, auditory discrimination, or language proficiency. The sample was not intended to be nationally representative, but was carefully selected to cover a wide range of community sizes, socio-economic indices and geographic regions. A total of approximately 10,500 children, or about 750 per form, were involved in the program.

The MRT was administered by classroom teachers with each class receiving only one form of the test. The teachers were instructed to work with

small groups of children so that they could become aware of any difficulties with marking, handling the test booklet, or similar problems. The items were read aloud; in cases where no reading was required (e.g. Visual Matching) the items were paced to make sure each child had a chance to respond to each item.

The booklets were machine scored and the data analyzed by computer. The cross-tabulations for each item (number of correct responses by subtest score and population group) and other data necessary for the bias study were provided as part of the output. Each item was then tested for bias using a 2 x r chi square technique, where there were two population groups and r score groups. It would be possible, though perhaps not desirable, to do the analysis with more than two population groups at a time, but only two groups in this study, Blacks and Whites, had a sufficient number of children on any one form to be analyzed. Population group membership was identified by each child's teacher. The value of r varied from item to item, because it was necessary to combine adjacent score groups, particularly at the extremes of the distribution, in order to get large enough expected frequencies in all cells. The number of score groups required varied with the difficulty of the item and the length of the subtest. When the probability of the obtained chi square value for an item fell below .30, it was recommended that the item be dropped from the pool.

Results

Using data from other parts of the item analysis program, seven subtest areas at each of the two levels were selected for inclusion in the final forms of the MRT. Therefore, only items in the 44 subtests in these areas were screened for bias. Excluding a number of items within these subtests which were dropped for other reasons, the chi square tests were performed on a final pool of

579 items. Of these, eighty, or about 14 percent, were termed biased by this procedure.

Table I summarizes the results by ability area. It can be seen that the majority of biased items fell into the Language area. Of these, 23 items, or 29 percent of all biased items, were from Quantitative subtests although the 110 Quantitative items made up only 19 percent of the total item pool. The fewest number of biased items came from the Visual area. In the Visual Matching subtests at Level II, only one out of every 45 items was called biased.

The content of the biased items was then reexamined to determine possible reasons for bias, but in most cases the cause of the bias was not immediately obvious. In the case of items from the School Language tests at Level II, however, the biased items tended to fall into a pattern. Of the 55 School Language items tests, ten involved some negative structures. For example, "Mark the thing that is unopened." or "Mark the picture which shows neither a cat nor a dog." Of the seven items found to be biased in these subtests, six involved the negative forms, while a seventh item involving negatives was only slightly above the cut-off point. This pattern seemed too strong to ignore even though the apparent direction of the bias was not consistent. For this reason, all ten items involving negatives were dropped from the item pool.

The chi square test treats a deviation from equality in either direction as equivalent, so in order to infer the direction of the bias, it was necessary to look at the performance of children at each score level in more detail. In doing so, four distinct patterns emerged.

1. The item was apparently biased against Blacks (Group A).
2. The item was apparently biased against Whites (Group B).

3. The item was probably not biased at all. If adjacent score categories were combined into still larger categories, the difference between groups tended to disappear.
4. For high scoring children the difference was in one direction and for low scoring children it was in the other.

Further examination of items displaying the fourth pattern, which will be called the "differential validity pattern," revealed that the point-biserial correlations between the item and the subtest total score for the two population groups were quite different. With these items a large proportion of high scoring children and a small proportion of low scoring children from one group got the item correct as would be expected. In the other group, however, the change in proportion from high to low scorers was relatively slight.

In order to illustrate the types of patterns which emerged, items were selected from one subtest, the Quantitative subtest from Booklet 8 Level II. The performance of the two population groups on this subtest is summarized on Table II, with results from the sample items shown in Table III. For each item, Table III provides the difficulty value (proportion of correct responses); the point-biserial correlation with the subtest total; the percent of correct responses in each score group, separately for the two population groups; and the chi square statistics. Item 3 shows a typical differential validity pattern. The distribution of correct responses across score groups is relatively flat for Group A, so that comparatively Group B scores much better than expected at the highest score level and less well at the lowest score level. Item 8 is an unbiased item presented for contrast with the others. Item 9 shows

that Group B does consistently less well than Group A except for a small difference in the opposite direction for the lowest scoring group. In item 10, Group A does consistently less well than Group B.

Discussion

Although the results of the bias study reported here probably do not warrant any generalizable conclusions, there are some observations which can be made. First, if a distribution were made of the differences between the item difficulty values of the two population groups for any one subtest, the items which were found to be clearly biased against one of these groups would tend to lie at the extremes of such distributions. The magnitude of the difference required for an item to appear at such extremes varies from subtest to subtest, but it is likely that the same items would have appeared using other techniques. It should be noted, however, that several subtests had no biased items and many had biased items at only one extreme.

In the case of items displaying the differential validity pattern, however, this was not so. Many of these items showed differences in difficulty values which were only moderate for their subtest. According to the definition of bias presented here, these items are biased if not clearly in favor of one group or another. Yet a method which uses difficulty value, even in some transformation or in relation to other items, as the sole variable reflecting bias is apt to miss these items.

It is possible that a large difference in point-biserial correlation or other indicator of internal consistency could be used as a separate screening device, but the question then arises of how large a difference would justify

dropping an item. In this study, all six items where the difference between the point-biserial r's exceeded .30 were found to be biased. However, for a number of items displaying the differential validity pattern, the differences were between .20 and .30, a range which also contained a number of items which were evidently not biased. The chi square test would appear to provide a meaningful criterion for determining when the departure between groups becomes significant.

Another observation concerns the number of score groups formed in doing the chi square tests. In this study, the number of groups was generally kept as high as the different distributions permitted. However, this procedure does not appear to be optimal. The results which seemed to be spurious occurred most often if the number of score groups was either very low or very high. On the basis of experience, four to six groups would seem to be most satisfactory if the data allow that many to be formed.

In conclusion, the chi square procedure described in this report appears to be a satisfactory technique for screening out items which are likely to be biased before the final construction of an ability or achievement measure.

References

Angoff, W.H. & Ford, S.F. Item-Race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Cleary, T.A. & Hilton, T.L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

TABLE I

Number of Biased Items by Area

Area		Biased Items	Total Items Tested
Language	N	49	287
	%	61	50
Visual	N	15	154
	%	19	27
Auditory	N	16	138
	%	20	24
Total	N	80	579
	%	100	100

TABLE I I

Quantitative Subtest
Level II Booklet 8
23 Items

	Population Group	
	A	B
N	103	615
x	9.12	13.01
sd	3.00	3.61
Q3	11.18	15.49
Med.	8.94	13.11
Q_1	6.88	10.47
Range	3-18	3-22

TABLE III

Examples of Items
Quantitative Subtest

Item	Population Group	Difficulty	Pt. Biserial	% of Correct Responses* Hi	Med. Hi	Med. Lo	Lo	X^2	df	Prob**
3	A	.29	.09	36	29	25	27	4.66	3	<.20
	B	.56	.48	70	30	32	17			
8	A	.50	.43	92	68	37	27	0.70	3	>.80
	B	.78	.48	93	72	50	25			
9	A	.50	.34	75	65	54	21	6.34	3	<.10
	B	.53	.33	65	43	29	28			
10	A	.40	.42	83	50	33	18	6.00	3	<.20
	B	.78	.42	94	68	61	32			

* Hi - scores of 12-22 for item 3
13-22 for items 8-10

Med-Hi - scores of 10-11 for item 3
10-12 for items 8-10

Med-Lo - scores of 8-9

Lo - scores of 3-7

** The probability that the item is not biased.